Vol.2013-SLP-97 No.6
2013/7/26

統計的言語モデルにおける確率的潜在意味解析の学習初期化手法の一検討

大島 寛史[1,a)]　川端 豪[1]

概要：確率的潜在意味解析（以下 PLSA ）は，文書の話題を反映した言語モデルを構築する手法である．PLSA は初期値として与えられた話題 unigram を基に自己組織化を行うが，統計学習の際の初期値に 0 が多く含まれるか否かでの言語モデルとしての PLSA の振る舞いの変化を捉える．結果として，初期値に 0 を多く含む場合は補正 Perplexity は低く抑えられ，学習も早い段階で収束するものの，初期値依存性が強く，また，未知語の割合が高まることが観測された．他方，0 を含まない場合には初期値依存性は弱まるものの，補正 Perplexity は 0 を含む場合と比べて数割程度増加し，収束までに必要な EM アルゴリズムのステップ数が大幅に増加した．

キーワード：PLSA，言語モデル，EM アルゴリズム

Consideration about the Sparseness of Parameter Initialization of PLSA Language Models

Abstract: PLSA (Probabilistic Latent Semantic Analysis) is a promising technology to reduce the perplexity for speech recognition systems. In this method, the topic structure is self-organized as the topic unigram vector. This paper describes parameter initialization methods taking the sparseness of topic vectors into account. The perplexity reduction experiments show that the sparse initialization of topic vectors enables the faster and more accurate topic cluster organization. However in this case, the ratio of unknown words increases and the dependency to initial data selection also increases.

Keywords: PLSA，language model，EM algorithm

1. はじめに

現在，大語彙連続音声認識の際に用いられる言語モデルとしては，N-gram モデルが一般的である [1]．N-gram モデルでは直前に出現する $N-1$ 単語を基に確率を推定するため，話題や話し方のスタイルなど，文脈全体から得られる特徴については，そのごく一部しか反映することはできない．そこで，それらの文脈情報を取り入れることで，言語モデルを現状よりも最適化することが可能となる．

話題やスタイルを言語モデルに反映させる手法には大きく分けて二種類ある．

一つは，認識結果から推定される話題を基にその話題に関連のある文書を収集し，それらを学習データとして言語モデルを再構築する手法である．この手法に関する研究としては，有効な検索クエリを構築することで，WWW から学習データを収集するというものがある [2], [3].

もう一つは，複数の話題が混在している学習データから話題推定を行うという手法である．マルコフモデルを用いて話題を制御する方法 [4] や MAP 推定を用いたタスク適応の研究 [5] などがなされている．この手法を用いる利点は，目標とする文書と類似した話題を持つ文書集合を用意せずに済むという点である．また，話題をモデル化する手段として確率的潜在意味解析 [6] がある．これは，認識対象の話題を判定し，単語の出現確率に重みをつけることで，言語モデルをタスクに適応させるものである．学習の最適化の方法や [7]，語彙分割に関する検討 [8]，話者適応による音声認識率の改善 [9]，WWW から得られる検索語重み

[1] 関西学院大学大学院
Kwansei Gakuin Uniersity
[a)] cqd79253@kwansei.ac.jp

ⓒ 2013 Information Processing Society of Japan

付けを用いた話題適応 [10] などの発展研究が存在する．しかしながら，その性能が統計学習の際に与える初期値に強く依存することなど，いくつかの問題点が存在している．

2. 単語の頻度分布に基づく話題モデル

文脈や話題などの言語的な情報を用いて発話される音声の文型や語彙を絞り込み，認識性能を上げることが，音声認識における言語モデルの重要な役割である．何らかの話題を設定し，想定される語彙を制限する手法において，話題の種類や，それに対応する語彙集合を人手で設定するのは困難である．この問題を解決するためには，それらの要素を自動的に決定するような手法が望まれる．

本節では，そのような手法の一つである確率的潜在意味解析の統計学習に用いる EM アルゴリズムにおいて，初期値の与え方，特に学習の際に与える初期値に含まれる 0 の有無による統計的言語モデルとしての確率的潜在意味解析の振る舞いを観察し，性能や初期値依存性の変化を捉える．

2.1 確率的潜在意味解析の枠組み

確率的潜在意味解析 (Probabilistic Latent Semantic Analysis ，以下 PLSA) [6] とは，学習データから得られる単語の出現頻度を基に，話題をモデル化する手法である．PLSA が k-means 法 [11], [12] などの話題をモデル化する手法と違っているのは，複数の話題が入り混じったような，複雑な話題に対しても効果を発揮するところである．PLSA では，内部に話題ごとにその特徴を反映した unigram モデル（単語出現確率ベクトル）を持ち，それらの話題 unigram を適切に混合することにより，目的の話題に対し最適化された unigram を得ることができる．(図 1)

PLSA における話題 h を反映した単語 w の出現頻度 $P(w|h)$ は以下の式 (1) で与えられる．

$$P(w|h) = \sum_{z \in Z} P(w|z)P(z|h) \quad (1)$$

ここで，$P(w|z)$ は話題 unigram z が単語 w に対して与える確率となる．他方，$P(w|h)$ は目的の話題 h に対して最適な話題 unigram の混合比である．

2.2 PLSA による話題集合および話題 unigram の学習形成

PLSA は内部パラメータとして話題 z における単語 w の出現頻度を表す $P(w|z)$ と，文書 d における話題 unigram の混合比を表している $P(z|d)$ を持つ．$P(z|d)$ は $P(z)$ 及び $P(d|z)$ からベイズの定理によって導き出すことができる．これらの値を用いて，学習データに含まれる 1 文書ごとの単語の出現回数を学習データとし，EM アルゴリズムにより反復学習を繰り返すことで，以下の式 (2) を最大化することで，尤度を最大化するような任意の数の話題 unigram を学習することで生成される．

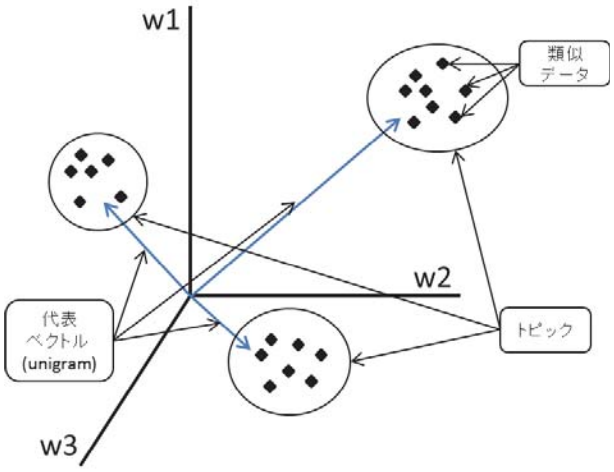

図 1 PLSA の概念図

$$l(\theta; N) = \sum_{w \in W} \sum_{d \in D} n(d, w) \log \sum_{z \in Z} P(w|z)P(z|d) \quad (2)$$

ここで $n(d, w)$ は文書 d における単語 w の出現回数を表す．

統計学習には Tempered EM アルゴリズム (以下 T-EM) という反復学習法を用いる．T-EM に用いる式は以下の式 (3) ∼ (6) となる．

E-Step:

$$P^{(k)}(z|d, w) = \frac{\{P^{(k)}(z)P^{(k)}(d|z)P^{(k)}(w|z)\}^{\beta}}{\sum_{z \in Z}\{P^{(k)}(z)P^{(k)}(d|z)P^{(k)}(w|z)\}^{\beta}} \quad (3)$$

M-Step:

$$P^{(k+1)}(w|z) = \frac{\sum_{d \in D} n(d, w)P^{(k)}(z|w, d)}{\sum_{w \in W}\{\sum_{d \in D} n(d, w)P^{(k)}(z|w, d)\}} \quad (4)$$

$$P^{(k+1)}(d|z) = \frac{\sum_{w \in W} n(d, w)P^{(k)}(z|w, d)}{\sum_{d \in D}\{\sum_{w \in W} n(d, w)P^{(k)}(z|w, d)\}} \quad (5)$$

$$P^{(k+1)}(z|d, w) = \frac{\sum_{w \in W}\sum_{d \in D} n(d, w)P^{(k)}(z|d, w)}{\sum_{w \in W}\sum_{d \in D} n(d, w)} \quad (6)$$

E-Step と M-Step を交互に繰り返すことで式 (2) を最大化するモデルを生成することができる．

2.3 PLSA を用いた話題同定

PLSA では話題に重み付けを行い，目的文書に即した話題の混合を行う．ここで目的文書とは，例えば，直前に発話された音声の認識結果，あるいは，今から認識しようとしている音声の認識候補である．これらの文書が PLSA の内包するどの話題にどれだけ属するかを，話題 unigram の混合比として推定する．これは話題 unigram の混合比を目的の文書に対し最尤推定により最適化することで行う．その際には学習時と同じく T-EM を用いる．T-EM に用いる式は以下の式 (7),(8) となる．

E-step:

$$P^{(k)}(z|h, w) = \frac{\{P(z)P^{(k)}(h|z)P(w|z)\}^{\beta}}{\sum_{z \in Z}\{P(z)P^{(k)}(h|z)P(w|z)\}^{\beta}} \quad (7)$$

ⓒ 2013 Information Processing Society of Japan

M-step:

$$P^{(k+1)}(h|z) = \frac{\sum_{w \in W} n(h,w) P^{(k)}(z|h,w)}{\sum_{z \in Z} \{\sum_{w \in W} n(h,w) P^{(k)}(z|h,w)\}} \quad (8)$$

2.4 アニーリング

T-EM においては，学習速度と局所最適解への落ち込みを防ぐ目的で，アニーリングと呼ばれる操作が行われる．通常の EM アルゴリズムとの違いとして，式（3）のように E-Step の際に右辺全体を β 乗 ($\beta \geq 0$) する．$\beta = 1.0$ の場合に通常の EM アルゴリズムと等しくなる．β が 1.0 より小さければ尤度関数が平滑化され，局所最適解への収束を防ぐ等の効果がある．

T-EM ではこの β を反復学習が進行するに連れて変化させていく．この β を変化させる手続きをアニーリングスケジュールという．アニーリングスケジュールは大きく分けて二種類ある．一つは β の初期値を 1.0 よりも小さい値から始め，徐々に増やしていくことで最終的に 1.0 にするもので，DAEM (Deterministic Annealing EM) と呼ばれる．これには，学習初期段階での局所最適解への収束を防ぐ効果がある．他方 β の初期値を 1.0 とし，学習が進むにつれて β を減らしていくもので，inverse annealing と呼ばれる．こちらには，学習を加速させ，また，過学習を防ぐ効果がある．本稿では学習の速度を早める目的から inverse annealing により学習を行うこととする．

3. PLSA 学習における初期値の検討

PLSA の学習に際して，はじめに話題数 L を与える．無作為に学習データより抽出した L 個の文書の単語 unigram を各話題 unigram の初期値として用いる．その際，初期値として与える値に 0 を多く含むスパースな unigram を用いる場合と，すべての要素に何らかの値を与える場合について，形成された PLSA による，補正 Perplexity の違いを比較する．

3.1 初期値としての 0

PLSA の学習の際に用いる式 (3) 及び (4) より，初期化の際に特定の z に対して $P^{(0)}(w|z) = 0$ となる w が存在するとき，k の値にかかわらず，$P^{(k)}(w|z) = 0$ となる．このことから，初期化において $P(w|z)$ に 0 を与える場合，PLSA は他の値を与える場合と比べて特別な振る舞いをすると考えられる．

仮に初期化の際に 0 が多く含まれた場合，0 を持つ部分に関しては計算の必要性がなくなるため，学習の際に必要な計算コストの削減が可能となる．しかしながら，抽出された文書の特徴が強く反映されるため，初期値依存性が高まることが懸念される．また，未知語率の上昇も考えられる．他方，初期化の際に 0 を含まない場合には，学習における収束が遅くなる，あるいは最適解に辿りつけないなどの問題点が考えられる．一方で，初期値依存性は大幅に改善されると考えられる．

そこで初期値に 0 を含む場合と含まない場合における PLSA の統計的言語モデルとしての性能と初期値依存性の比較を行った．

表 1　PLSA の学習条件

学習データ	CSJ
文書数	972
語彙数	約 10000
潜在モデル数	50
EM 反復回数	100
アニーリングスケジュール	inverse annealing
β 初期値	1.0
β 終端値	0.8
β 更新回数	4

3.2 実験条件

(1) 式における $P(w|z)$ を，以下の 3 通りの方法で初期化する．

1. 乱数により無作為に L 個の文書を選び，それらに含まれる単語頻度の分布に基づいて L 個の話題 unigram を初期化する．乱数を変えて 5 通り行う．

2a. 1. で得られた話題 unigram の各要素を一様にフロアリングする．具体的には，1. で得られた $P_1(w|z)$ から以下の式によって求める

$$P_{2a}(w|z) = p \cdot P_1(w|z) + (1-p) \cdot \frac{1}{n(W)} \quad (9)$$

この時, $n(W)$ は語彙数を表す．また，今回は $p = 10^{-6}$ とした．1. で作成した乱数を変えた 5 通りの初期値をベースに各々フロアリングを行う．

2b. 1. で得られた値を全学習データを用いて得た unigram により平滑化する．具体的には, 1. で得られた $P_1(w|z)$ から以下の式によって求める

$$P_{2b}(w|z) = p \cdot P_1(w|z) + (1-p) \cdot uni_{all}(w) \quad (10)$$

この時，$uni_{all}(w)$ は学習データすべてを使って構築した unigram における単語 w の出現確率を表す．また，今回は $p = 10^{-6}$ とした．1. で作成した乱数を変えた 5 通りの初期値をベースに各々フロアリングを行う．

このとき，1. の手法によって得られた初期値には 0 が多く含まれるのに対して，2. の方法で得られた初期値には 0 が含まれることはない．また，2a. と 2b. を比較することで平滑化の手法による実験結果への影響を観測することができると考えられる．

PLSA の学習条件を表 1 に示す．学習データには日本語話し言葉コーパス (以下 CSJ) に含まれる実講演データ 987

表 2　PLSA の適応条件

EM 反復回数	60
アニーリングスケジュール	inverse annealing
β 初期値	1.0
β 終端値	0.9
β 更新回数	2

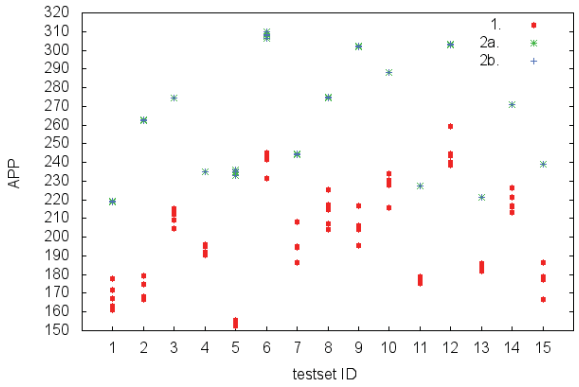

図 2　初期化法による PLSA の性能の散布図

講演のうち，評価用データ 15 講演分を除いた 972 講演分のデータを用いた．語彙は学習データに含まれる単語のうち，10 回以上出現した約 1 万単語とした．また，潜在モデル数は 50，学習時の EM アルゴリズムの反復回数は 100 回とした．T-EM の β 更新のためのアニーリングスケジュールには inverse annealing を用い，最終的な β の値が 0.8 になるように 5 段階に更新を行った．

評価の際の PLSA の適応条件を表 2 に示す．テストセットには CSJ に含まれる実講演データ 987 講演から無作為に抽出した 15 講演を用いた．評価の際の EM アルゴリズムの反復回数は 60 とし，また，適応時のアニーリングスケジュールは過学習を防止するため inverse annealing を用い，2 段階に更新を行った．評価尺度にはテストセットに対する補正 Perplexity を用いた．補正 Perplexity の計算式は次式（11）になる．

$$APP = \{P(w_1, w_2 ... w_n) \cdot m^{-o}\}^{-\frac{1}{n}} \tag{11}$$

このとき，$P(w_1, w_2 ... w_n)$ は単語列 $w_1, w_2 ... w_n$ が生成される確率を，O は未知語の数を，m は未知語の種類数をそれぞれ表す．

$P(d|z)$ と $P(z)$ については式（12）及び式（13）により初期化を行った．

$$P(d|z) = \frac{1}{n(D)} \tag{12}$$

$$P(z) = \frac{1}{L} \tag{13}$$

この時 $n(D)$ は学習データに含まれる文書数を，L は潜在モデル数をそれぞれ表している．

3.3　実験結果

初期化法による PLSA の性能の散布図を図 2 に示す．グラフの横軸は用いたテストセットを，縦軸は補正 Perplexity の値を表す．各テストセットに対し，1. による 5 種の結果（赤丸），2a. による 5 種の結果（緑＊），2b. による 5 種の結果（青＋）がプロットされている．ただし，緑と青についてはほとんど縮退して一つの点に見える．

この図より 1. の手法がもっとも補正 Perplexity を低く抑えることができているのがわかる．その一方で 1. においては乱数のシードにより最大で 15% 程度の補正 Perprexity の大きな変動が見られた．このことから，初期値に 0 を多く含む 1. の手法では初期値依存性が強く現れることがわかる．一方，図中の 2a. と 2b. では乱数のシードによる差がほぼ見られない．この原因としては 2 つの可能性が考えられる．一つは初期化の際に 0 を含ませないことで初期値依存性が弱くなる可能性．そしてもうひとつは，初期化の際に 0 を含ませないことで学習に必要な EM アルゴリズムの反復回数が増える，または，最適解に辿りつけなくなっている可能性である．また，この実験においては学習の際の EM アルゴリズムの反復回数を定数としていたことから，もし初期値によって EM アルゴリズムによる学習の進行速度に変化がある場合，その差が性能差として現れてしまった可能性がある．

3.4　初期値の性質の違いによる学習時の振る舞い

初期値の性質の違いによる学習時の振る舞いを観測するため，3.3 と同条件で EM 反復回数のみを 200 にし，10 回毎の補正 Perprexity の変化から，学習時の振る舞いについて観察を行った．

その結果を図 3 および図 4 に示す．図の縦軸は補正 Perprexity を，横軸は EM アルゴリズムの学習における反復回数となっている．それぞれの図において，下側の点が 1. の手法により初期化した場合を，上側の点が 2b. の手法により初期化を行った場合を表す．

テストセットの 15 文書における学習時の振る舞いを観察した結果，大きくこれらの 2 種に分けられることがわかった．図 3 のような振る舞いをするものは，学習が進むにつれてゆるやかに減少していっているのが見て取れる．これは初期値に 0 を含まないことで，学習に必要な反復回数が増加していると捉えることができる．他方，図 4 のような振る舞いをするものは，学習を繰り返しても性能が向上することはなく，早い段階で局所解に収束してしまっていることが伺える．

4.　おわりに

本稿では，確率的潜在意味解析において，初期値の与え方，特に学習の際に与える初期値に 0 が含まれるかどうかによるによる統計的言語モデルとしての確率的潜在意味解

ⓒ 2013 Information Processing Society of Japan

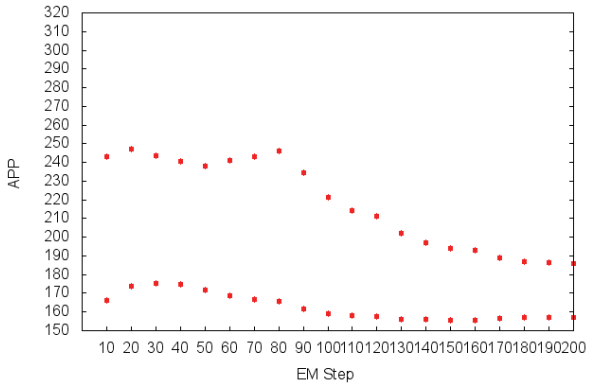

図 3 初期化法による PLSA の学習進行の違い 1

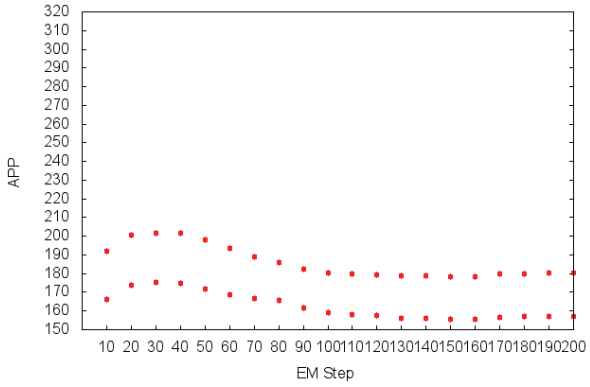

図 4 初期化法による PLSA の学習進行の違い 2

析の振る舞いの変化を観察した．

結果として，初期値に 0 を多く含むことを認めた場合，学習が早く終わり，また，補正 Perprexity においても良い性能がでる結果となった．しかしながら，乱数のシードによる性能のブレが大きいといった欠点がある．また，この初期化法においては，未知語率が高くなることが確認されており，実際に音声認識に持ち込んだ場合の性能には不安が残る．他方，初期値に 0 を含まない場合には，学習の速度が遅くなることがわかった．また，場合によっては局所最適解への収束が見られた．こちらの手法では，未知語率は 0 を含む場合の半分以下に抑えられていた．

参考文献

[1] 北研二：確率的言語モデル，東京大学出版会 (1999).
[2] 梶浦泰智，鈴木基之，伊藤彰則，牧野正三: WWW を利用した言語モデル教師なしタスク適応における有効検索クエリ決定法, 情処研報, 2006-SLP-64(2006).
[3] 増村亮，伊藤仁，伊藤彰則，牧野正三: WWW を利用した言語モデル適応のための検索クエリ構成の検討, 情処研報，Vol.2009-SLP-76 No. 10(2009).
[4] 長野雄, 鈴木基之, 牧野正三：HMM を用いた複数 n-gram モデルによる言語モデルの構築，情処論, Vol.J43 , No.7, pp.2075-2081(2002).
[5] 政瀧浩和, 匂坂芳典, 久木和也, 河原達也：MAP 推定を用いた N-gram 言語モデルのタスク適応，信学技報, SP96-103(1997).
[6] Thomas Hofmann : *Probabilistic Latent Semantic Analysis*, Uncertainity in Artificial Intelligence (1999).
[7] Daniel Gildea, Thomas Hofmann : *TOPIC-BASED LANGUAGE MODELS USING EM*, EuroSpeech’99, pp.2167-2170(1999).
[8] 栗山直人, 鈴木基之, 伊藤彰則, 牧野正三：PLSA 言語モデルの学習最適化と語彙分割に関する検討，情処研報, 2006-SLP-60(2006).
[9] 秋田祐哉, 河原達也：話題と話者に関する PLSA に基づく言語モデル適応，情処研報, 2003-SLP-49(2003).
[10] 宮崎将隆：WWW から得られる Term Frequency 情報に基づく PLSA 言語モデル，情処研報, 2011-SLP-85, No.14(2011).
[11] 宮本定明：クラスター分析入門，森北出版（1999）.
[12] 岸田和明：文書クラスタリングの技法，*Library and Information Science*, Vol. 49, pp. 33?75, 2003.

ⓒ 2013 Information Processing Society of Japan